ORIGINAL RESEARCH

Prediction of Diabetes using Logistic Regression, Classification, and Regression Tree

Jagdish D. Powar,*[1,2] Rajesh Dase,[2] Deepak Bhosle[3]

Abstract

Background: Diabetes is a chronic disease that occurs because of an imbalance in blood sugar levels. The prevalence of diabetes mellitus is rising rapidly worldwide. More than 50% of people are unaware of undiagnosed diabetes. The use of artificial intelligence is increasing in the healthcare industry for the management, prediction, and diagnosis of diseases. Classification and regression tree (CART) is a predictive model in artificial intelligence that can be used for early prediction with better accuracy. The logistic regression model is a supervised machine learning model, used for predicting categorical variables.

Objective: To design a prediction model for type 2 diabetes using classification and regression tree (CART) and logistic regression.

Methodology: In the proposed research, classification and regression tree (CART) and logistic regression model are used for prediction. The Pima Indian Diabetes Database (PIDD) was used to develop the prediction model. The performance of the algorithms is evaluated based on precision, accuracy of classification.

Result: We found that the accuracy of the logistic regression was 78.32% and that of the CART was 80.75%. The CART algorithm sensitivity was 81.95%, whereas the logistic regression sensitivity was 88.93%. For CART analysis, specificity was 80.15 and 56.92% for logistic regression.

Conclusion: The study underscores the significance of predicting diabetes for prevention, employing logistic regression and CART analysis to identify crucial risk factors like glucose, insulin, age, and BMI. CART analysis showed superiority in accuracy, sensitivity, and specificity over logistic regression.

Keywords: Artificial intelligence, classification and regression tree, Prediction, Diabetes

Journal of Research in Medical & Interpathy Sciences. 2(1);2024

Introduction

Diabetes is a complex and chronic medical condition that affects millions of people worldwide. It is characterized by the body's inability to effectively regulate blood sugar levels, leading to consistently high levels of glucose in the blood. The number and prevalence of people with diabetes is rapidly increasing.[1] This condition can lead to a range of complications, including heart disease, kidney damage, and vision problems. Managing diabetes requires a combination of medication, lifestyle modifications, and regular monitoring of blood sugar levels.

Diabetes is a growing health concern worldwide, and India is no exception. With a population of more than 1.3 billion people, India has the world's second highest number of diabetic adults. The prevalence of diabetes has been on the rise in recent years, posing significant challenges to healthcare systems globally. The weighted prevalence of diabetes reported in the ICMR-INDIAB-17 study was 11·4% and that of prediabetes was 15·3%.[2] Estimates in 2019 showed that 77 million individuals had diabetes in India, and this number is expected to rise to over 134 million by 2045[2]. Approximately 57% of individuals have undiagnosed diabetes mellitus. According to the International Diabetes Federation, approximately 537 million adults (20-79 years) are living with diabetes across the globe.[3] An estimated 83.8% of all cases of undiagnosed diabetes mellitus occur in low- and middle-income countries.[3] This high prevalence of diabetes in India can be attributed to various factors, including genetic predisposition, sedentary lifestyles, unhealthy dietary habits, and a lack of awareness regarding proper diabetes management.

Diabetes is one of the most prevalent and rapidly spreading diseases in many nations, and individuals are attempting to avoid it at an early stage by anticipating the signs of diabetes using a variety of techniques.[1-4]

[1]Department of Community Medicine, SMBT, IMSRC, Nashik, Maharashtra, India

[2]Department of Community Medicine, MGM, Medical College & Hospital, Aurangabad, Maharashtra, India

[3]Department of Pharmacology, MGM, Medical College & Hospital, Aurangabad, Maharashtra, India

Corresponding Author: Jagdish D. Powar, Department of Community Medicine, SMBT, IMSRC, Nashik, Maharashtra, India. Email: jdpstat1479@gmail.com

Conflict of Interest: None

Source of Funding: None

Artificial intelligence (AI) is now an attractive technology in the healthcare industry. AI is an effective tool for identifying diagnosed and undiagnosed patients with various diseases by considering family history and other biological parameters with previous records. Predicting such a chronic disease at an early stage of development is required to prevent diabetes and its associated health problems. In the proposed research, the development of diabetes prediction models is performed using a logistic regression model and a classification and regression tree.

Methodology

The proposed methodology involves systematic implementation of logistic regression and classification and regression tree (CART) to diagnose diabetes. The Pima Indian Diabetes Dataset (PIDD) was retrieved from the UCI website and used for the prediction of diabetes and model development.[5] The PIDD historical data consists of 8 attributes on 768 individuals, including features such as pregnancies, age, body mass index (BMI), glucose levels, blood pressure, skin thickness, insulin, diabetes pedigree function, and a binary outcome variable indicating diabetes status as 0 for non-diabetic and 1 for diabetic.

Data Cleaning

Cleaning of the dataset was performed because it contains missing and extreme values. Individuals with missing or extreme values were excluded from the study. The variables of glucose, blood pressure, skin thickness, insulin, and BMI were observed with missing values. The variable insulin contains maximum missing values (i.e., 375 and skin thickness contains 227, '0' values. Individuals with missing or 0 values and extreme values were excluded from the dataset to remove inconsistencies. Data from 377 individuals were excluded from the PIDD, and data from 392 individuals were used for model development and prediction of diabetes using logistic regression and the CART classification technique. After exclusion, data from 130 diabetic and 262 non-diabetic patients were used for model development. Logistic regression model predicts the probability of diabetes. It is a linear model used for binary classification tasks. CART creates a tree structure based on feature splits that maximize information gain or minimize impurity. CART is a non-linear, tree-based model that recursively partitions data based on features to make predictions. In this study, logistic regression, classification, and regression tree were used for training data and predicting the risk of diabetes. These models were then evaluated using various metrics, including accuracy, sensitivity, specificity, precision, and the ROC-AUC score. Minitab and Jamovi statistical software were used for prediction and model development.

Result

Figure 1 presents the relative importance of the predictor variables. Out of 8 attributes, we found that glucose, insulin, age, and BMI were relatively the most important risk factors in the development of diabetes (Figure 1).

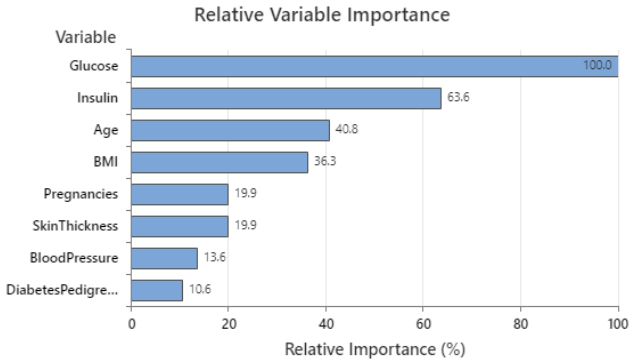

Figure 1: Relative importance of variable

Table 1: Output of the logistic regression and CART

Predictors	β-Coefficient	95% CI for OR	p-value	Odds Ratio
Constant	-10.04	--	0.0000	0.00
Glucose	0.04	(1.01, 1.05)	0.0000	1.04
Blood Pressure	0.00	(0.97, 1.02)	0.9040	1.00
Skin Thickness	0.01	(0.97, 1.04)	0.5110	1.01
Insulin	0.00	(0.99, 1.00)	0.5280	1.00
BMI	0.07	(1.02, 1.13)	0.0100	1.07
Age	0.03	(0.99, 1.08)	0.0650	1.03
Diabetes Pedigree Function	1.14	(1.35, 7.23)	0.0080	3.13
Pregnancies	0.08	(0.97, 1.21)	0.1380	1.09

Table 1 shows the odds ratio and β-coefficient of logistic regression. The β-coefficient in logistic regression was highly significant (*p-value* < 0.01) in terms of glucose, BMI, and diabetes pedigree function. The odds ratio of all predictors was nearly equal to 1 except for diabetes pedigree function, where the odds ratio was 3.13.

Figure 2 shows a six-nodal optimal tree diagram of the CART analysis using the Gini method. Glucose, insulin, and age are important predictors of diabetes. If a patient with glucose >127.5 and age > 23.5 patients will be predicted as diabetic. If a patient with glucose ≤ 127.5, insulin >143.5, and age > 28.5 years, they will be diagnosed as diabetic.
Figure 3 presents the ROC curve from CART analysis.

Table 2 shows the predictive measures of logistic regression and the CART algorithm. We observed the accuracy of CART was 80.75% and accuracy for logistic regression was 78.32%. The sensitivity of the CART algorithm was 81.95%, whereas the sensitivity for logistic regression was

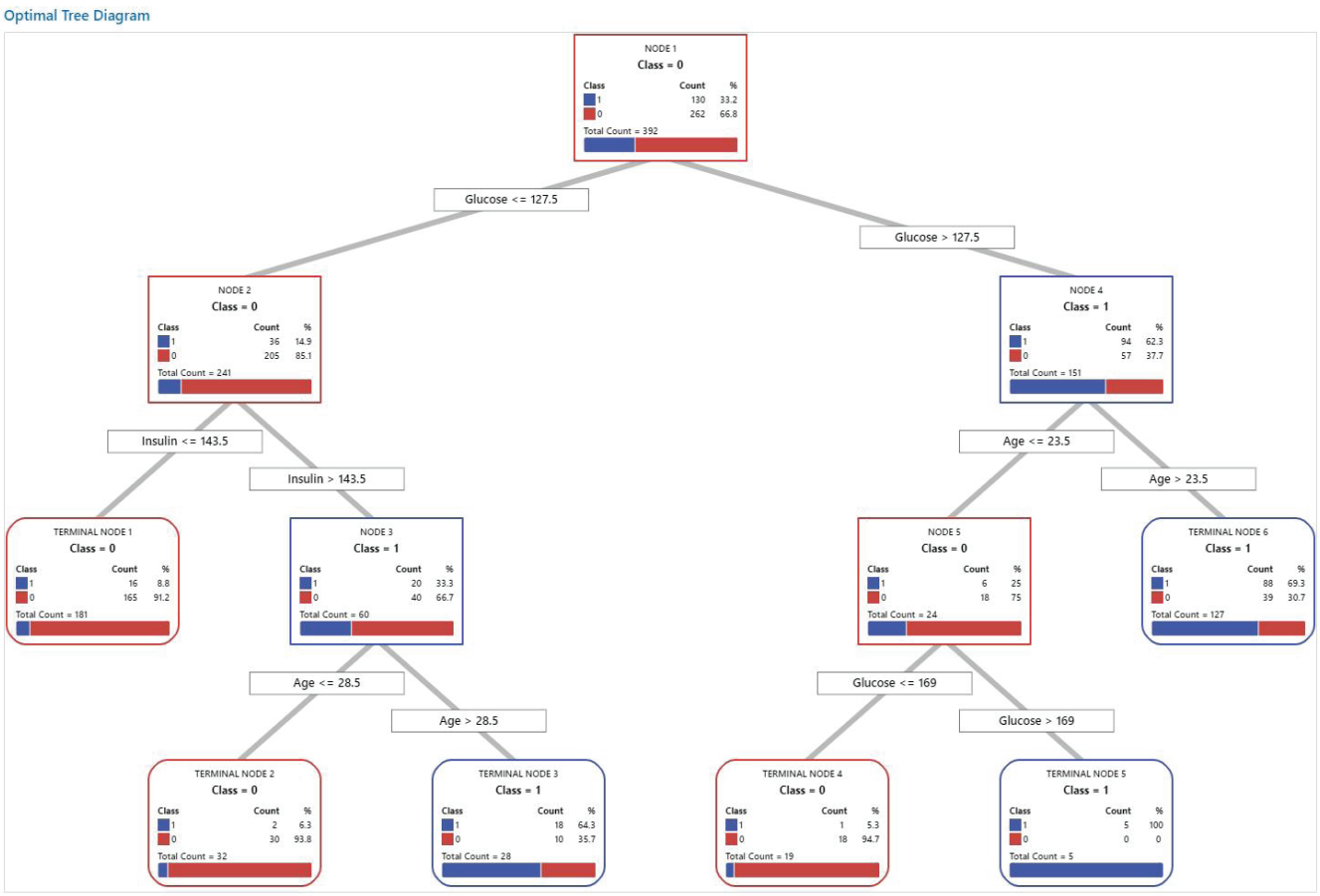

Figure 2: Tree diagram

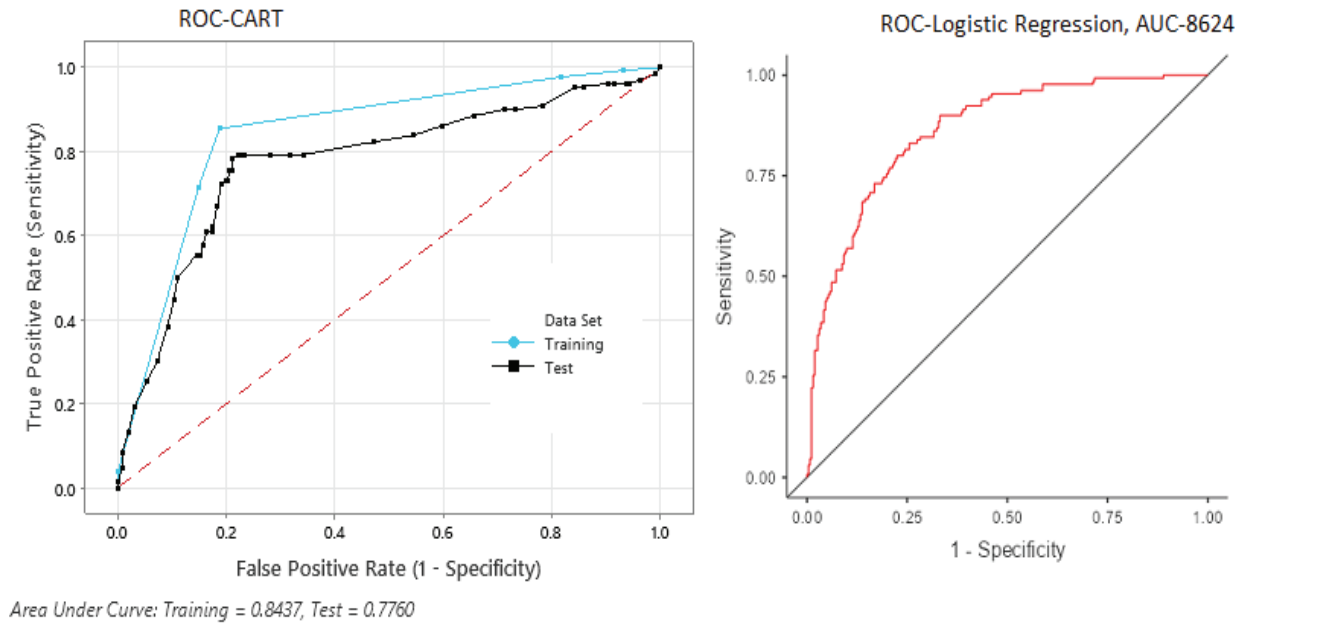

Figure 3: ROC curve from CART

Table 2: Predictive measures of CART and logistic regression

Predictive Measures	CART Analysis	Logistic Regression
Accuracy	80.75%	78.32%
Sensitivity	81.95%	88.93%
Specificity	80.15%	56.92%
AUC	0.8437	0.8624

88.93%. Specificity for CART analysis was 80.15% and it was 56.92%. The area under the curve (AUC) for logistic regression was 0.8624, whereas the AUC in CART was 0.8098.

Discussion

We developed a logistic regression model and a CART tree model and found that the accuracy of CART is greater than that of the logistic regression model. The sensitivity of logistic regression is greater than that of CART, whereas specificity, i.e., the false positive rate of CART, is greater than 80.15%. We found that glucose, insulin, age, and BMI are the most relative important risk factors in the development of diabetes. The β-coefficient of glucose, BMI, and diabetes pedigree function in the logistic regression model. Authors *Raghad Sehly* and *Mohammad Mezher* performed comparative analysis of the classification models for PIDD. They found that the accuracy of logistic regression was 76.95% and that of CART was 69.52%[6]. M. Seera *et al*. (2013) measured the accuracy of the CART model using the Pima Indian Diabetes Dataset. The accuracy of the CART model was 70.03%.[7] Researcher Neha Tigga found accuracy for logistic regression PIDD 74.4 and 85.7% on the data collected by the authors.[8] Another study predicted diabetes using machine learning algorithms.[9] On the dataset, various machine learning techniques were used by the authors and the highest accuracy 96% for the logistic regression algorithm. P. Suresh Kumar and Pranavi S used random forest, support vector machine, k-nearest neighbor, CART, and LDA algorithms for disease prediction and found an accuracy of 93.1% for CART analysis[10]. A case study in Dandipora district to estimate the prevalence of diabetes and predict diabetes at an early stage using machine learning[11] found that the accuracy of logistic regression was 69%. In a study authors performed diabetes prediction using machine learning on PIDD data and found the glucose, insulin, BMI, and age are the important risk features.[12] They observed the highest accuracy of 84% for the random forest algorithm. Xue-Hui Meng *et al*. compared three data mining models for predicting diabetes or prediabetes by risk factors logistic regression, an artificial neural network, and a C5 decision tree.[13] Their logistic regression model achieved a classification accuracy of 76.13%, the artificial neural network model achieved a classification accuracy of 73.23%, and the decision tree (C5.0) achieved a classification accuracy of 77.87%.[13]

Conclusion

The study underscores the importance of predicting diabetes for effective prevention strategies. By leveraging logistic regression and CART analysis, the research identified crucial predictors of diabetes onset. Interestingly, CART analysis emerged as superior in terms of accuracy, sensitivity, and specificity compared to logistic regression. The study highlights glucose, insulin, age, and BMI as pivotal risk factors for diabetes development, emphasizing the significance of monitoring these variables for disease control. Moving forward, these findings provide valuable insights for researchers aiming to refine prediction algorithms utilizing logistic regression and CART methodologies, ultimately aiding in the proactive management and prevention of diabetes on a global scale. Both the CART (Classification and Regression Trees) and logistic regression algorithms demonstrate reasonably high accuracy in predicting the outcome of interest. The CART algorithm outperformed logistic regression slightly, with an accuracy of 80.75% compared to 78.32%. However, when considering sensitivity, which reflects the ability of the models to correctly identify true positives, logistic regression performed better at 88.93% compared to CART's 81.95%. On the other hand, CART showed higher specificity at 80.15%, indicating its better ability to correctly identify true negatives compared to logistic regression, which had a specificity of 56.92%. These findings suggest that while logistic regression may be more sensitive in capturing true positive cases, CART may offer better performance in identifying true negatives.

Reference

1. Guariguata L, Whiting DR, Beagley J, Linnenkamp U, Hambleton I, Cho NH, et al. Global estimates of diabetes prevalence in adults for 2013 and projections for 2035 for the IDF Diabetes Atlas. Diabetes Res Clin Pract 2013. http://dx.doi.org/10.1016/j.diabres.2013.11.002.
2. Anjana RM, Unnikrishnan R, Deepa M, Pradeepa R, et al. Metabolic non-communicable disease health report of India: the ICMR-INDIAB national cross-sectional study (ICMR-INDIAB-17). Lancet Diabetes Endocrinol. 2023 Jul;11(7):474-489. doi: 10.1016/S2213-8587(23)00119-5. Epub 2023 Jun 7. PMID: 37301218.
3. Beagley J, Guariguata L, Weil C, Motala AA. Global estimates of undiagnosed diabetes in adults. Diabetes Res Clin Pract. 2014 Feb;103(2):150-60. doi: 10.1016/j.diabres.2013.11.001. Epub 2013 Dec 1. PMID: 24300018.
4. Pima Indians Diabetes Database. Data. World. Available online at: https://data.world/data-society/pima-indians-diabetes-database (retrieved 03 Jun 2023).
5. Sehly R and Mezher M. Performance Impact of Genetic Operators in a Hybrid GA-KNN Algorithm. International Journal of Advanced Computer Science and Applications2020;11(11): 476-87.
6. M. Seera, C. P. Lim, S. C. Tan, and C. K. Loo, "A hybrid FAM- CART model and its application to medical data classification," Neural Computing and Applications, vol.

26, no. 8, pp. 1799-1811, 2015.
7. Tigga NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science 2020;167: 706–716.
8. Mujumdar A, Vaidehi V. Diabetes Prediction using Machine Learning Algorithms. Procedia Computer Science 2019; 165: 292–9.
9. Kumar PS, Pranavi S. Performance analysis of machine learning algorithms on diabetes dataset using big data analytics. International Conference on Infocom Technologies and Unmanned Systems 2017: 508-513.
10. Bhat SS, Selvam V, Ansari GA, Ansari MD, Rahman MH. Prevalence and early prediction of diabetes using Machine Learning in North Kashmir: A case study of district Bandipora. Computational Intelligence and Neuroscience 2022:1-12.
11. Dutta D, Paul D and Ghosh P. Analysing Feature Importances for Diabetes Prediction using Machine Learning. IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference 2018: 924-928.
12. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung Journal of Medical Sciences 2013; 29:93-99.